



US006263334B1

(12) **United States Patent**
Fayyad et al.

(10) Patent No.: **US 6,263,334 B1**
(45) Date of Patent: **Jul. 17, 2001**

(54) **DENSITY-BASED INDEXING METHOD FOR EFFICIENT EXECUTION OF HIGH DIMENSIONAL NEAREST-NEIGHBOR QUERIES ON LARGE DATABASES**

4-6, 1997, pp. 599-608.*

(List continued on next page.)

(75) Inventors: **Usama Fayyad**, Mercer Island, WA (US); **Kristin P. Bennett**, Troy, NY (US); **Dan Geiger**, Tivon (IL)

Primary Examiner—Kim Vu

Assistant Examiner—Shahid Alam

(74) Attorney, Agent, or Firm—Watts, Hoffmann, Fisher & Heinke, Co., L.P.A.

(73) Assignee: **Microsoft Corporation**, Redmond, WA (US)

(57) **ABSTRACT**

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

Method and apparatus for efficiently performing nearest neighbor queries on a database of records wherein each record has a large number of attributes by automatically extracting a multidimensional index from the data. The method is based on first obtaining a statistical model of the content of the data in the form of a probability density function. This density is then used to decide how data should be reorganized on disk for efficient nearest neighbor queries. At query time, the model decides the order in which data should be scanned. It also provides the means for evaluating the probability of correctness of the answer found so far in the partial scan of data determined by the model. In this invention a clustering process is performed on the database to produce multiple data clusters. Each cluster is characterized by a cluster model. The set of clusters represent a probability density function in the form of a mixture model. A new database of records is built having an augmented record format that contains the original record attributes and an additional record attribute containing a cluster number for each record based on the clustering step. The cluster model uses a probability density function for each cluster so that the process of augmenting the attributes of each record is accomplished by evaluating each record's probability with respect to each cluster. Once the augmented records are used to build a database the augmented attribute is used as an index into the database so that nearest neighbor query analysis can be very efficiently conducted using an indexed look up process. As the database is queried, the probability density function is used to determine the order clusters or database pages are scanned. The probability density function is also used to determine when scanning can stop because the nearest neighbor has been found with high probability.

(21) Appl. No.: **09/189,229**

(22) Filed: **Nov. 11, 1998**

(51) Int. Cl.⁷ **G06F 17/30**

(52) U.S. Cl. **707/5; 704/9; 706/50; 707/6; 707/100; 707/102; 709/202**

(58) Field of Search **707/5, 6, 100, 707/101, 200, 102; 704/9, 222, 233, 245; 706/45, 50; 705/1, 10, 27; 702/179; 709/202**

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,325,298 * 6/1994 Gallant 704/9
5,590,242 * 12/1996 Juang et al. 704/245
5,781,704 * 7/1998 Rossmo 706/45
5,787,422 7/1998 Tukey et al. .
5,790,426 * 8/1998 Robinson 702/179
5,832,182 * 11/1998 Zhang et al. 706/50
5,884,282 * 3/1999 Robinson 705/27

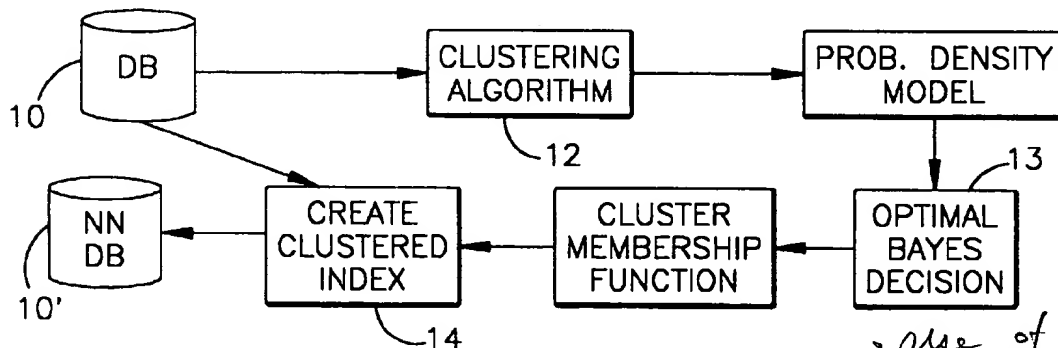
FOREIGN PATENT DOCUMENTS

797 161 A2 9/1997 (EP) .

OTHER PUBLICATIONS

Kleinberg, Jon M., "Two Algorithms for Nearest-Neighbor Search in High Dimensions", Proceedings of the twenty-ninth annual ACM symposium on Theory of Computing, May

39 Claims, 6 Drawing Sheets



• use of VA-file
(col. 2)
• nearest-neighbor
queries

OTHER PUBLICATIONS

- Nabaa, Nassib et al., "Derivation and Analytic Evaluation of an Equivalence Relation Clustering Algorithm", IEEE Transactions on Systems, Man, And Cybernetics—Part B: Cybernetics, vol. 29, Issue 6, Dec. 1999, pp. 908–912.*
- Roussopoulos, Nick et al., "Nearest Neighbor Queries", Proceedings of the 1995 ACM SIGMOD international conference on Management of Data, May 22–25, 1995, pp. 71–79.*
- Shimoji, Shunichi et al., "Data Clustering with Entropical Scheduling", IEEE International Conference on Neural Networks; IEEE World congress on Computational Intelligence, Jun. 27–Jul. 2 1994, vol.: 4, pp. 2423–2428.*
- S. Berchtold et al., "A Cost Model For Nearest Neighbor Search in High-Dimensional Data Space". In proceedings 14th International Conference on Data Engineering. (1998).
- S. Berchtold et al., "Fast Nearest Neighbor Search in High-Dimensional Space". In ACM PODS Symposium on Principles of Database Systems, Tucson, Arizona, (1997).
- S. Berchtold et al., The Pyramid-Technique: "Towards Breaking The Curse of Dimensionality", pp. 142–153, Seattle, WA. In Proceedings of ACM SIGMOD International Conference On Management of Data, (1998).
- S. Berchtold et al., High-Dimensional Index Structures: "Database Support For Next Decade's Application". Tutorial notes: ACM SIGMOD-98 Conference on Management Conference On Management of Data, Seattle, WA. pp. 1–65, (1998).
- S. Berchtold et al., The X-Tree: "An Index Structure For High-Dimensional Data". In Proceedings of the 22nd Conference on Very Large Databases", pp. 28–39 Bombay, India, (1996).
- K. Beyer et al., "When Is nearest Neighbor Meaningful"? In Proceedings of the 7th International Conference On Database Theory (ICDT) pp. 1–19, Jerusalem, Israel, (1999) (1998).
- C. Faloutsos et al., The TV-Tree: "An Index Structure For High-Dimensional Data". VLDB Journal 3(4): pp.181–210, (1994).
- C. Faloutsos et al., Fastmap: "A Fast Algorithm For Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets". In Proceedings of ACM SIGMOD International Conference On Management of Data, San Jose, pp. 1–25, (1995).
- R. Farebrother. Algorithm as 204: "The Distribution of a Positive Linear Combination of Chi-Square Random Variables". Applied Statistics, 32(3):332–337, (1983).
- N. Katayama et al., The SR-Tree: "An Index Structure For High-Dimensional Nearest Neighbor Queries". In Proceedings of ACM SIGMOD International Conference On Management of Data, pp. 1–12, Tucson, Arizona, (1997).
- T. Seidl et al., "Optimal Multi-Step K-Nearest Neighbor Search". In Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 154–165 (1998).
- Shim et al., The ϵ -K-D-B Tree: "A Fast Index Structure For High-Dimensional Similarity Joins". In the 13th International Conference On Data Engineering. (unknown).
- White et al., "Similarity Indexing With The SS-Tree". In Proceedings of the 12th International Conference on Data Engineering, New Orleans, pp. 516–523, (1996).
- Indyk et al., Approximate Nearest Neighbors: "Towards Removing The Curse of Dimensionality", pp. 1–13 (1998).
- C. Bishop, Neural Networks for Pattern Recognition: "Bayes' Theorem", Clarendon Press. Oxford pp. 17–23 (1995).
- C. Bishop, Neural Networks For Pattern Recognition: "The Normal Distribution", Clarendon Press. Oxford, pp. 34–38 (1995).
- C. Bishop, Neural Networks For Pattern Recognition: "Maximum Likelihood" Clarendon Press. Oxford, pp. 39–42 (1995).
- C. Bishop, Neural Networks For Pattern Recognition: "Density Estimation In General" Clarendon Press. Oxford pp. 51–55 (1995).
- C. Bishop, Neural Networks For Pattern Recognition: "Mixture Models" Clarendon Press. Oxford pp. 59–72 (1995).
- A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood From Incomplete Data Via The EM Algorithm". Journal of The Royal Statistical Society, Series B, 39(1): 1–38, (1977).
- K. Fukunaga, Statistical Pattern Recognition: "Nearest Mean Reclassification Algorithm" (k-mean): Chapter 11 pp. 515–523, Academic Press (1990).
- E. Forgy, "Cluster Analysis of Multivariate Date: Efficiency vs. Interpretability of Classifications", Biometrics 21:768. (1965).
- T. Zhang et al., BIRCH: "A New Data Clustering Algorithm and its Applications, Data Mining and Knowledge Discovery" 1(2). (1997).

* cited by examiner

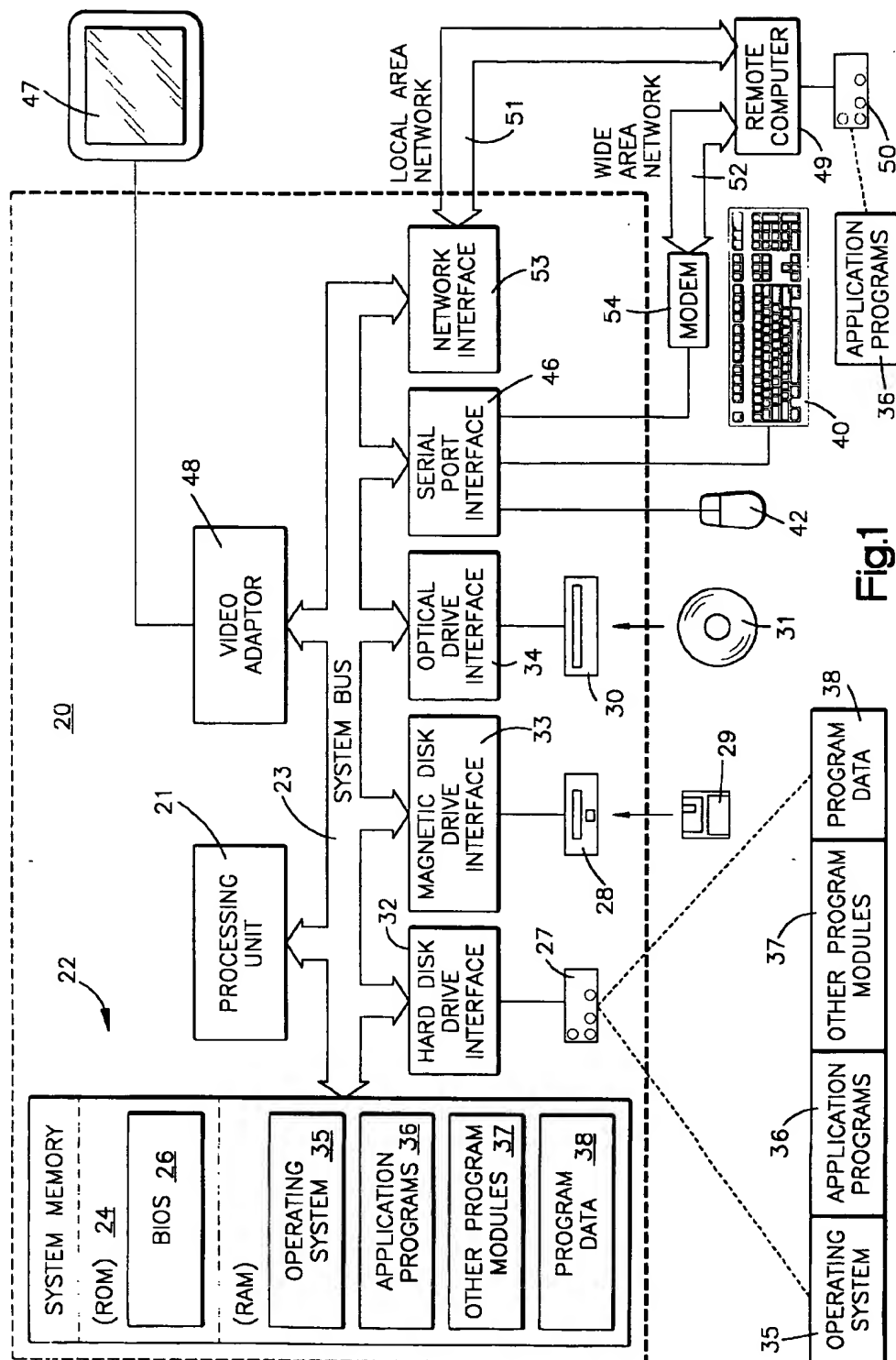
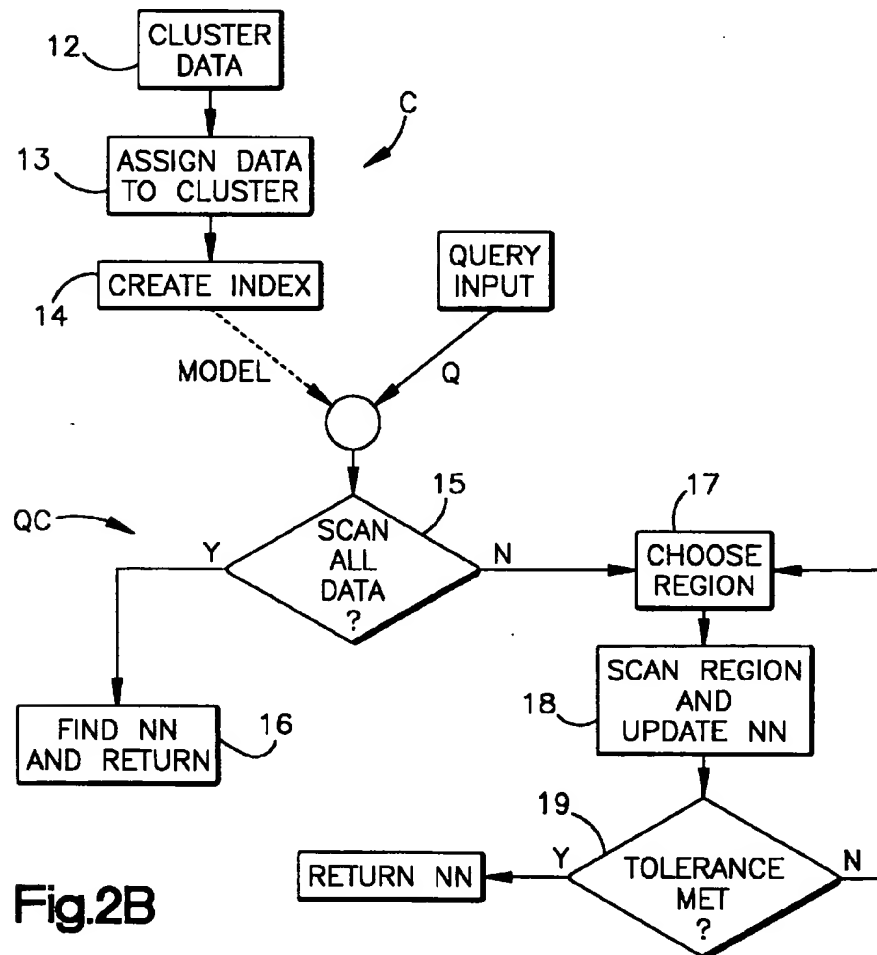
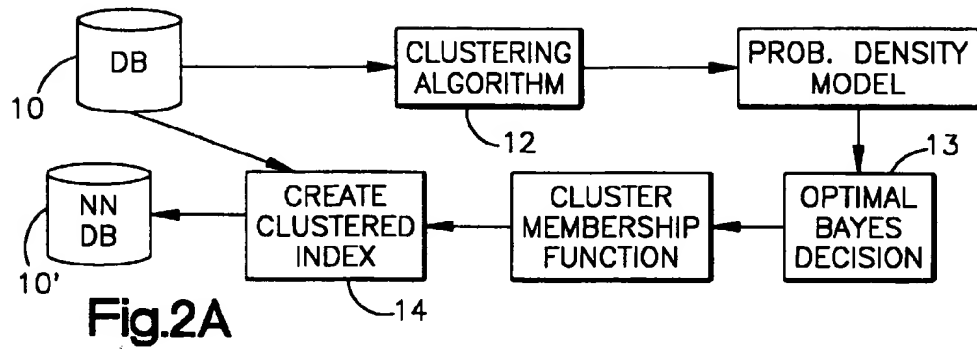
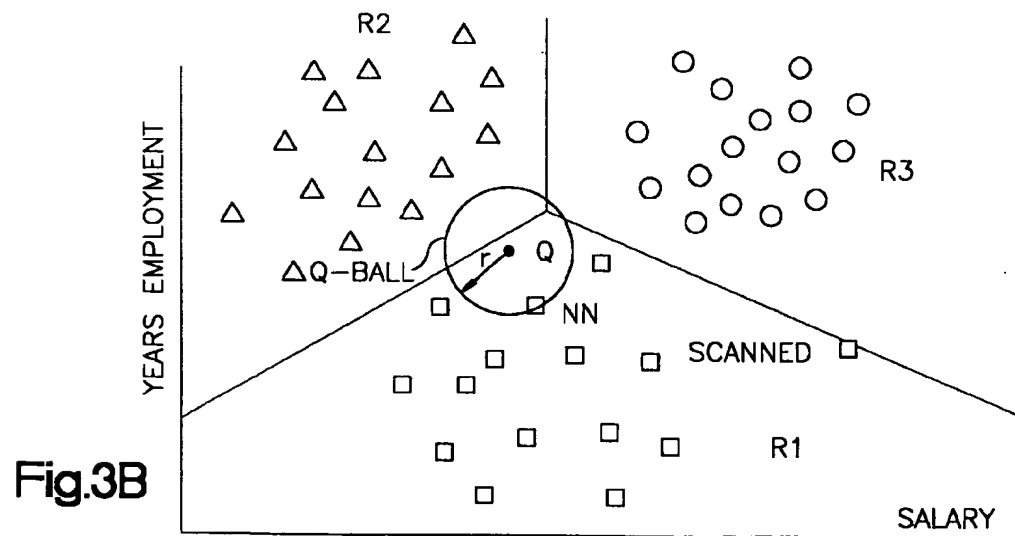
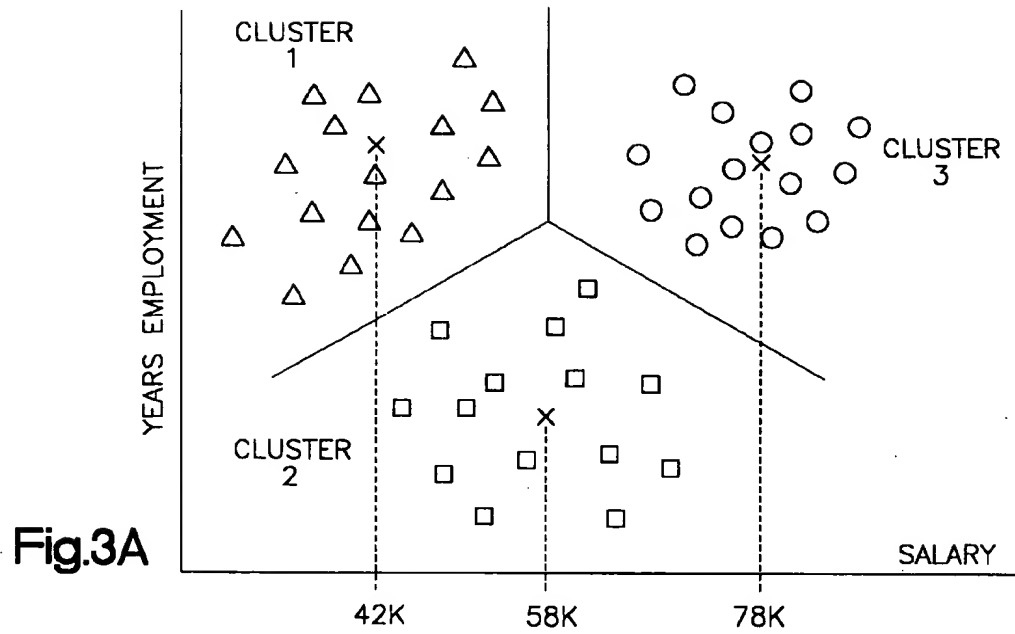


Fig. 1





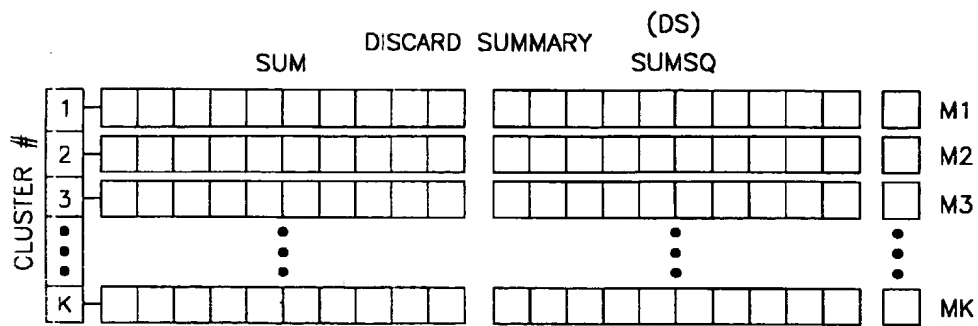


Fig. 4A

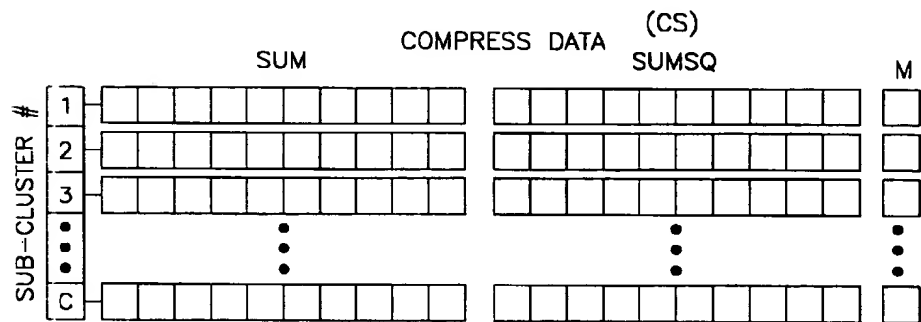


Fig. 4B

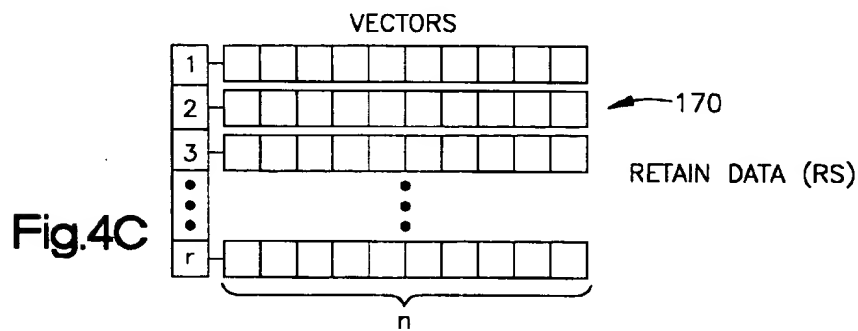


Fig. 4C

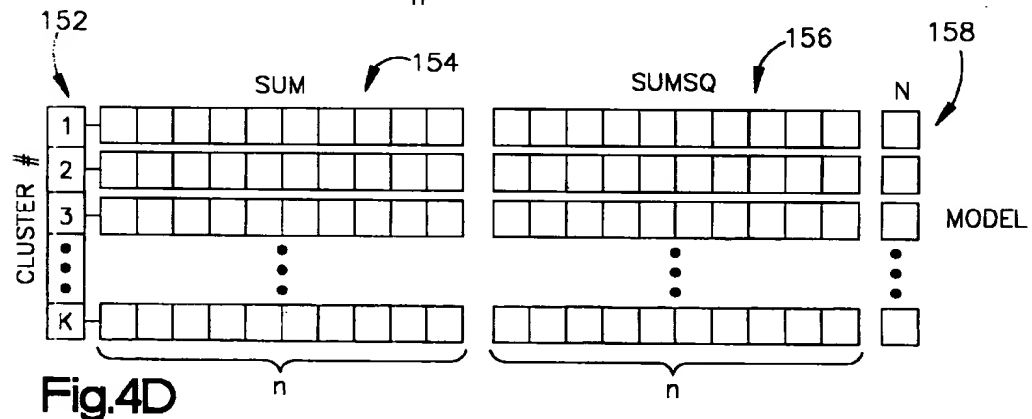
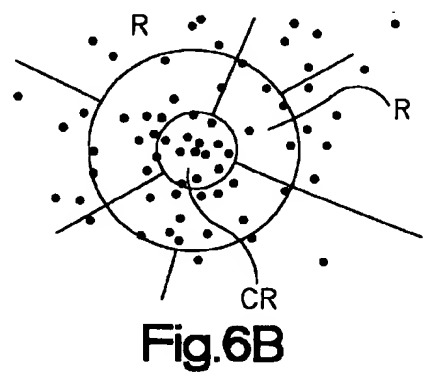
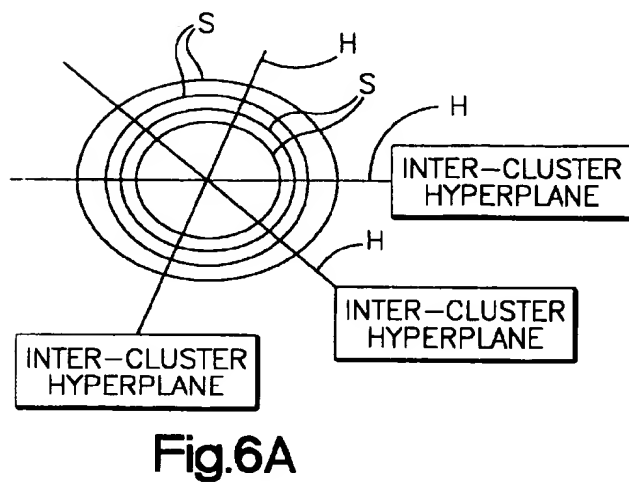
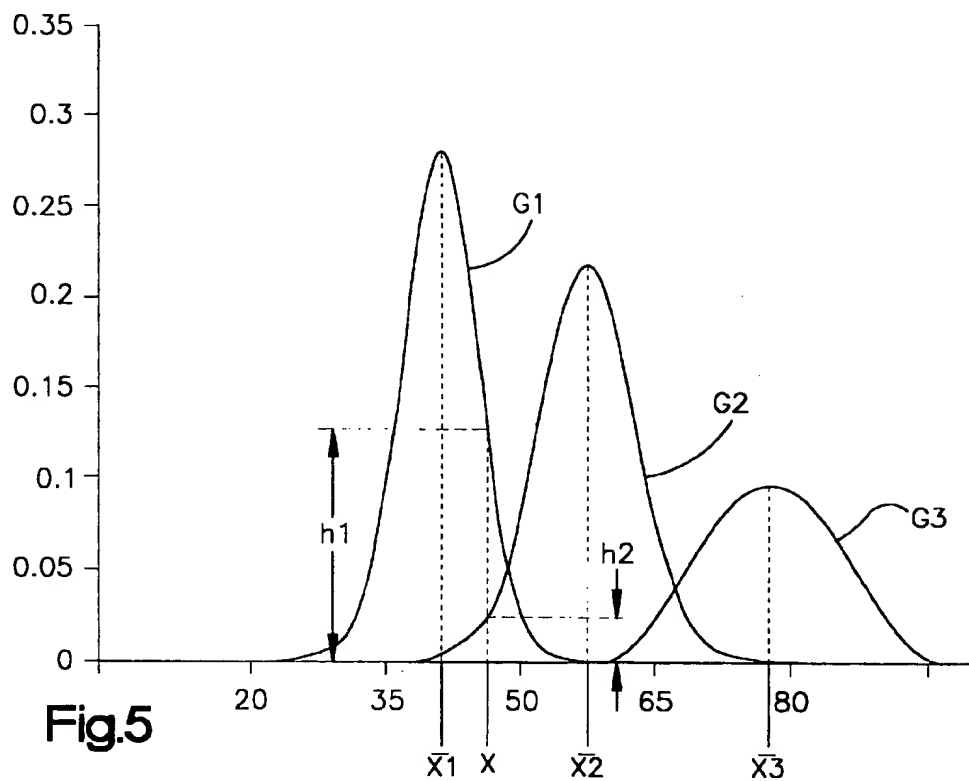
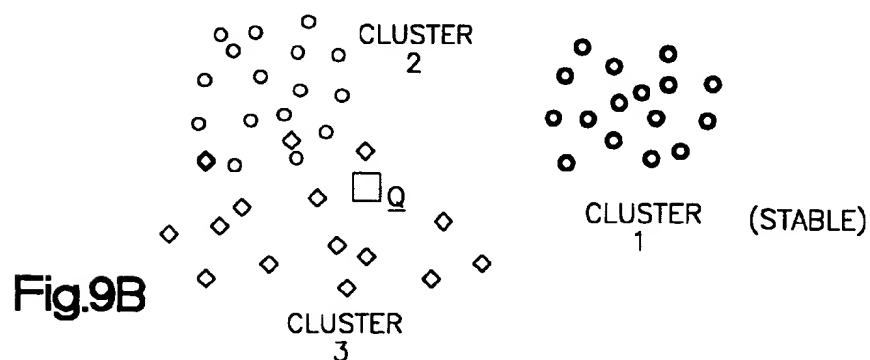
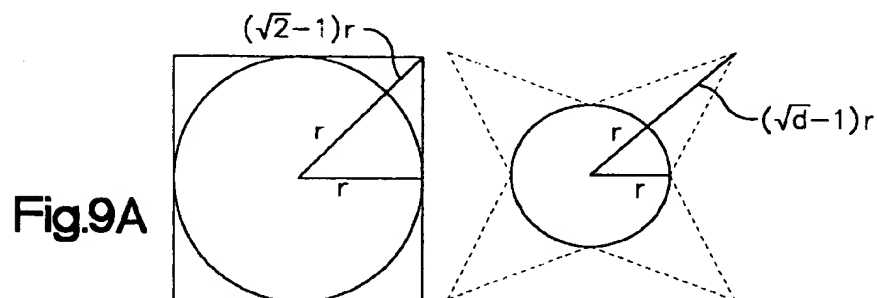
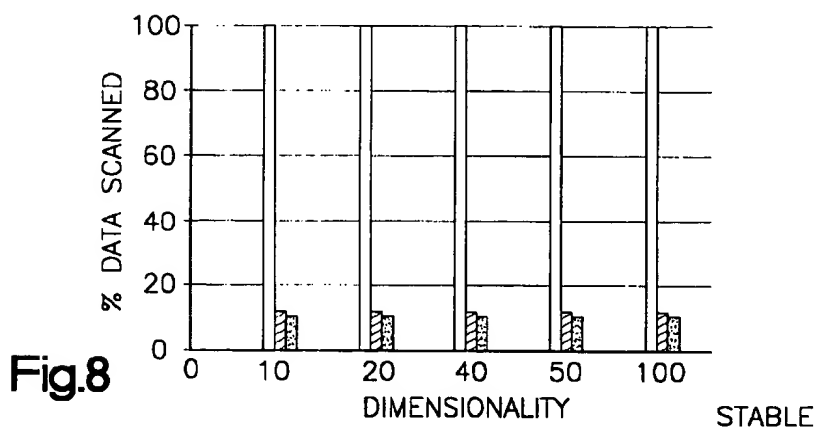
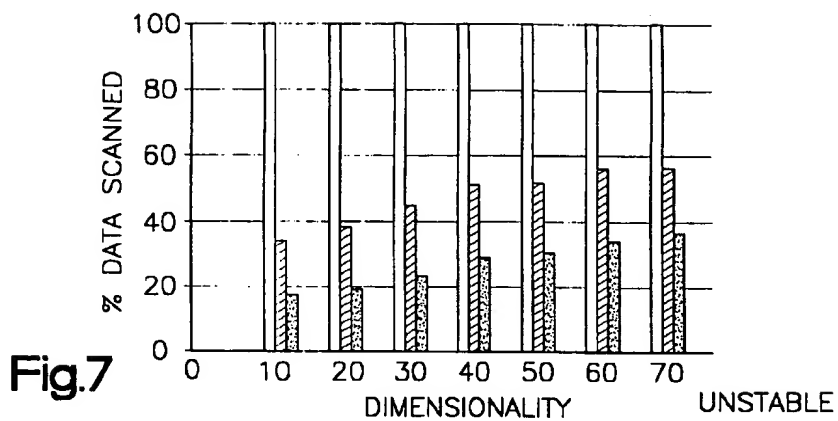


Fig. 4D





a storage medium have multiple data attributes that are described or summarized by a probability function. A nearest neighbor query is performed by assigning an index for each of the records based upon the probability function and then efficiently performing the nearest neighbor query.

In a typical use of the invention, the data is stored with the help of a database management system and the probability function is determined by performing a clustering of the data in the database. The results of the clustering are then used to create a clustered-index structure for answering nearest neighbor queries on the data stored in the database. The clustering identifies groups in the data consisting of elements that are generally "more similar" to each other than elements in other groups. The clustering builds a statistical model of the data. This model is used to determine how the data should be partitioned into pages and also determines the order in which the data clusters or pages should be scanned. The model also determines when scanning can stop because the nearest neighbor has been found with very high-probability.

Preliminary results on data consisting of mixtures of Gaussian distributions shows that if one knows what the model is, then one can indeed scale to large dimensions and use the clusters effectively as an index. Tests have been conducted with dimensions of 500 and higher. This assumes that the data meets certain "stability" conditions that insure that the clusters are not overlapping in space. These conditions are important because they enable a database design utility to decide whether the indexing method of this invention is likely to be useful for a given database. It is also useful at run-time by providing the query optimizer component of the database system with information it needs to decide the tradeoff between executing an index scan or simply doing a fast sequential scan.

An exemplary embodiment of the invention evaluates data records contained within a database wherein each record has multiple data attributes. A new database of records is then built having an augmented record format that contains the original record attributes and an additional record attribute containing a cluster number for each record based on the clustering model. Each of the records that are assigned to a given cluster can then be easily accessed by building an index on the augmented data record. The process of clustering and then building an index on the record of the augmented data set allows for efficient nearest neighbor searching of the database.

This and other objects, advantages and features of the invention will become better understood from the detailed description of an exemplary embodiment of the present invention which is described in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a schematic depiction of a computer system for use in practicing the present invention;

FIG. 2A is a schematic depiction of a data mining system constructed in accordance with an exemplary embodiment of the present invention;

FIG. 2B is a flow chart documenting the processing steps of an exemplary embodiment of the invention;

FIGS. 3A and 3B schematically illustrate data clusters;

FIGS. 4A-4D depict data structures used during representative clustering processes suitable for use in practicing the present invention;

FIG. 5 is a depiction in one dimension of three Gaussians corresponding to the three data clusters depicted in FIGS. 3A and 3B;

FIGS. 6A and 6B depict subdivisions of a cluster into segments in accordance with an alternative embodiment of the invention;

FIGS. 7 and 8 are tables illustrating nearest neighbor query results achieved through practice of an exemplary embodiment of the invention;

FIG. 9A is a two dimensional illustration of why high dimensional data records make nearest neighbor inquiries difficult; and

FIG. 9B is a depiction of three data clusters showing one stable cluster and two unstable clusters.

DETAILED DESCRIPTION OF EXEMPLARY EMBODIMENT OF THE INVENTION

The present invention has particular utility for use in answering queries based on probabilistic analysis of data contained in a database 10 (FIG. 2A). Practice of the invention identifies data partitions most likely to contain relevant data and eliminates regions unlikely to contain relevant data points. The database 10 will typically have many records stored on multiple, possibly distributed storage devices. Each record in the database 10 has many attributes or fields. A representative database might include age, income, number of years of employment, vested pension benefits etc. A data mining engine implemented in software running on a computer 20 (FIG. 1) accesses data stored on the database and answers queries.

An indexing process depicted in FIG. 2B includes the step of producing 12 a cluster model. Most preferably this model provides a best-fit mixture-of-Gaussians used in creating a probability density function for the data. Once the cluster model has been created, an optimal Bayes decision step 13 is used to assign each data point from the database 10 into a cluster. Finally the data is sorted by their cluster assignment and used to index 14 the database or optionally created an augmented database 10' (FIG. 2A) having an additional attribute for storing the cluster number to which the data point is assigned. As an optional step in the indexing process one can ask whether it makes sense to index based upon cluster number. If the data in the database produces unstable clusters as that term is defined below, then a nearest neighbor query using probability information may make little sense and the indexing on cluster number will not be conducted.

One use of the clustering model derived from the database 10 to answer nearest neighbor queries concerning the data records in the database. FIG. 2B depicts a query analysis component QC of the invention. Although both the clustering component C and the query component QC are depicted in FIG. 2B, it is appreciated that the clustering can be performed independently of the query. The query analysis component QC finds with high probability a nearest neighbor (NN) of a query point Q presented as an input to the query analysis component. The nearest neighbor of Q is then found in one of two ways. A decision step 15 determines whether a complete scan of the database is more efficient than a probabilistic search for the nearest neighbor (NN). If the complete scan is more efficient, the scan is performed 16 and the nearest neighbor identified. If not, a region is chosen 17 based on the query point and that region is scanned 18 to determine the nearest neighbor within the region. Once the nearest neighbor (NN) in the first identified region is found, a test is conducted 19 to determine if the nearest neighbor has been determined with a prescribed tolerance or degree of certainty. If the prescribed tolerance is not achieved a branch is take to identify additional regions to check for a nearest

neighbor or neighbors. Eventually the nearest neighbor or neighbors are found with acceptable certainty and the results are output from the query analysis component QC.

To illustrate the process of finding a nearest neighbor outlined in FIG. 2B consider the data depicted in FIGS. 3A and 3B. FIG. 3A is a two dimensional depiction showing a small sampling of data points extracted from the database 10. Such a depiction could be derived from a database having records of the format shown in Table 1:

TABLE 1

EmployeeID	Age	Salary	Years Employed	Vested Pension	Other Attributes
XXX-XX-XXXX	46	39K	15	100K	—
YYY-YY-YYYY	40	59K	4	0K	—
QQQ-QQ-QQQQ	57	88K	23	550K	—

The two dimensions that are plotted in FIG. 3A are years of employment and salary in thousands of dollars. One can visually determine that the data in FIG. 3A is lumped or clustered together into three clusters Cluster1, Cluster2, and Cluster3. FIG. 3B illustrates the same data points depicted in FIG. 3A and also illustrates an added data point or data record designated Q. A standard question one might ask of the data mining system 11 would be what is the nearest neighbor (NN) to Q in the database 10? To answer this question in an efficient manner that does not require a complete scan of the entire database 10, the invention utilizes knowledge obtained from a clustering of the data in the database.

Database Clustering

One process for performing the clustering step 12 of the data stored in the database 10 suitable for use by the clustering component uses a K-means clustering technique that is described in co-pending United States patent application entitled "A scalable method for K-means clustering of large Databases" that was filed in the United States Patent and Trademark Office on Mar. 17, 1998 under application Ser. No. 09/042,540 now, U.S. Pat. No. 6,012,058, and which is assigned to the assignee of the present application and is also incorporated herein by reference.

A second clustering process suitable for use by the clustering component 12 uses a so-called Expectation-Maximization (EM) analysis procedure. E-M clustering is described in an article entitled "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society B, vol 39, pp. 1-38 (1977). The EM process estimates the parameters of a model iteratively, starting from an initial estimate. Each iteration consists of an Expectation step, which finds a distribution for unobserved data (the cluster labels), given the known values for the observed data. Co-pending patent application entitled "A Scalable System for Expectation Maximization Clustering of Large Databases" filed May 22, 1998 under application Ser. No. 09/083,906 describes an E-M clustering procedure. This application is assigned to the assignee of the present invention and the disclosure of this patent application is incorporated herein by reference.

In an expectation maximization (EM) clustering analysis, rather than harshly assigning each data point in FIG. 3A to a cluster and then calculating the mean or average of that cluster, each data point has a probability or weighting factor that describes its degree of membership in each of the K clusters that characterize the data. For the EM analysis used in conjunction with an exemplary embodiment of the present invention, one associates a Gaussian distribution of data

about the centroid of each of the K clusters in FIG. 3A. EM is preferred over K-Means since EM produces a more valid statistical model of the data. However, Clustering can be done using any other clustering method, and then the cluster centers can be parametrized by fitting a gaussian on each center and estimating a covariance matrix from the data. EM gives us a fully parameterized model, and hence is the presently the preferred procedure.

Consider the one dimensional depiction shown in FIG. 5. The three Gaussians G1, G2, G3 represent three clusters that have centroids or means X1, X2, X3 in the salary attribute of 42K, 58K, and 78K dollars per year. The compactness of the data within a cluster is generally indicated by the shape of the Gaussian and the average value of the cluster is given by the mean. Now consider the data point identified on the salary axis as the point "X" of a data record having a salary of \$45,000. The data point 'belongs' to all three clusters identified by the Gaussians. This data point 'belongs' to the Gaussian G2 with a weighting factor proportional to h2 (probability density value) that is given by the vertical distance from the horizontal axis to the curve G2. This same data point X 'belongs' to the cluster characterized by the Gaussian G1 with a weighting factor proportional to h1 given by the vertical distance from the horizontal axis to the Gaussian G1. The point 'X' belongs to the third cluster characterized by the Gaussian G3 with a negligible weighting factor. One can say that the data point X belongs fractionally to the two clusters G1, G2. The weighting factor of its membership to G1 is given by $h1/(h1+h2+H_{rest})$; similarly it belongs to G2 with weight $h2/(h1+h2+H_{rest})$. Hrest is the sum of the heights of the curves for all other clusters (Gaussians). Since the height in other clusters is negligible one can think of a "fraction" of the case belonging to cluster 1 (represented by G1) while the rest belongs to cluster 2 (represented by G2). For example, if $h1=0.13$ and $h2=0.03$, then $0.13/(0.13+0.03)=0.8$ of the case belongs to cluster 1, while 0.2 of it belongs to cluster 2.

The invention disclosed in the above referenced two co-pending patent applications to Fayyad et al brings data from the database 10 into a computer memory 22 (FIG. 1) and the clustering component 12 creates an output model 14 from that data. The clustering model 14 provided by the clustering component 12 will typically fit in the memory of a personal computer.

FIGS. 4A-4D illustrate data structures used by the K-means and EM clustering procedures disclosed in the aforementioned patent applications to Fayyad et al. The data structures of FIGS. 4A-4C are used by the clustering component 12 to build the clustering model 14 stored in a data structure of FIG. 4D. Briefly, the component 12 gathers data from the database 10 and brings it into a memory region that stores vectors of the data in the structure 170 of FIG. 4C. As the data is evaluated it is either summarized in the data structure 160 of FIG. 4A or used to generate sub-clusters that are stored in the data structure 165 of FIG. 4B. Once a stopping criteria that is used to judge the sufficiency of the clustering has been achieved, the resultant model is stored in a data structure such as the model data structure of FIG. 4D.

Probability Function

Each of K clusters in the model (FIG. 4D) is represented or summarized as a multivariate gaussian having a probability density function:

computer-readable media provide nonvolatile storage of computer readable instructions, data structures, program modules and other data for the computer 20. Although the exemplary environment described herein employs a hard disk, a removable magnetic disk 29 and a removable optical disk 31, it should be appreciated by those skilled in the art that other types of computer readable media which can store data that is accessible by a computer, such as magnetic cassettes, flash memory cards, digital video disks, Bernoulli cartridges, random access memories (RAMs), read only memories (ROM), and the like, may also be used in the exemplary operating environment.

A number of program modules may be stored on the hard disk, magnetic disk 29, optical disk 31, ROM 24 or RAM 25, including an operating system 35, one or more application programs 36, other program modules 37, and program data 38. A user may enter commands and information into the computer 20 through input devices such as a keyboard 40 and pointing device 42. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 21 through a serial port interface 46 that is coupled to the system bus, but may be connected by other interfaces, such as a parallel port, game port or a universal serial bus (USB). A monitor 47 or other type of display device is also connected to the system bus 23 via an interface, such as a video adapter 48. In addition to the monitor, personal computers typically include other peripheral output devices (not shown), such as speakers and printers.

The computer 20 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 49. The remote computer 49 may be another personal computer, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 20, although only a memory storage device 50 has been illustrated in FIG. 1. The logical connections depicted in FIG. 1 include a local area network (LAN) 51 and a wide area network (WAN) 52. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

When used in a LAN networking environment, the computer 20 is connected to the local network 51 through a network interface or adapter 53. When used in a WAN networking environment, the computer 20 typically includes a modem 54 or other means for establishing communications over the wide area network 52, such as the Internet. The modem 54, which may be internal or external, is connected to the system bus 23 via the serial port interface 46. In a networked environment, program modules depicted relative to the computer 20, or portions thereof, may be stored in the remote memory storage device. It will be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be used.

While the present invention has been described with a degree of particularity, it is the intent that the invention include all modifications and alterations from the disclosed implementations falling within the spirit or scope of the appended claims.

We claim:

1. A method for evaluating data records contained within a database wherein each record has multiple data attributes; the method comprising the steps of:

- a) clustering the data records contained in the database into multiple data clusters wherein each of the multiple data clusters is characterized by a cluster model; and
- b) building a new database of records having an augmented record format that contains the original record attributes and an additional record attribute containing a cluster identifier for each record based on the clustering step.

2. The method of claim 1 wherein the cluster model includes a) a number of data records associated with that cluster, b) centroids for each attribute of the cluster model and c) a spread for each attribute of the cluster model.

3. The method of claim 1 additionally comprising the step of indexing the records in the database on the additional record attribute.

4. The method of claim 1 additionally comprising the step of finding a nearest neighbor of a query data record by evaluating database records indexed by means of the cluster identifiers found during the clustering step.

5. The method of claim 4 wherein the step of finding the nearest neighbor is performed by evaluating a probability estimate based upon a cluster model for the clusters that is created during the clustering step.

6. The method of claim 5 wherein the step of finding the nearest neighbor is performed by scanning the database records indexed by cluster identifiers having a greatest probability of containing a nearest neighbor.

7. The method of claim 6 wherein the step of scanning database records is performed for data records indexed on multiple cluster identifiers so long as a probability of finding a nearest neighbor within a cluster exceeds a threshold.

8. The method of claim 1 wherein the clustering process is performed using a scalable clustering process wherein a portion of the database is brought into a rapid access memory prior to clustering and then a portion of the data brought into the rapid access memory is summarized to allow other data records to be brought into memory for further clustering analysis.

9. The method of claim 1 wherein the number of data attributes is greater than 10.

10. The method of claim 4 wherein the step of finding the nearest neighbor is based upon quadratic distance metric between the query data record and the database records indexed by the cluster identifier.

11. A method for evaluating data records stored on a storage medium wherein each record has multiple data attributes that have been characterized by a probability function found by clustering the data in the database to produce a clustering model; said method comprising the steps of assigning a cluster number for each of the records on the storage medium based upon the probability function, and finding a nearest neighbor from the data records for a query record by scanning data records from more than one based upon the probability function of the cluster model used to assign cluster numbers to the data records.

12. The method of claim 11 wherein records having the same index are written to a single file on the storage medium.

13. The method of claim 11 wherein the records are stored by a database management system and the index is used to

19

form a record attribute of records stored within a database maintained on the storage medium.

14. The method of claim 11 wherein the clustering model for the database defines a probability function that is a mixture of Gaussians.

15. The method of claim 14 wherein the assigning of a cluster number comprises a Bayes decision for assigning each data record to a data cluster associated with one of said Gaussians.

16. The method of claim 11 wherein further comprising the step of subdividing data records within each cluster into cluster subcomponents and additionally comprising the step of finding a nearest neighbor of a query data by scanning records from a cluster subcomponent.

17. The method of claim 11 wherein the step of determining a nearest neighbor comprises the step determining a distance between the query data record and data records accessed in the database by means of the database index.

18. The method of claim 11 wherein a multiple number of data records are stored as a set of nearest neighbors.

19. The method of claim 11 wherein the nearest neighbor determination is based on a distance determination between a query record and a data record that comprises a quadratic normal form of distance.

20. The method of claim 11 wherein the storage medium stores a database of data records and further comprises a step of finding a nearest neighbor to a query record from the data records of the database, said step of finding the nearest neighbor comprising the step of choosing between a sequential scan of the database to find the nearest neighbor or searching for the nearest neighbor using the index derived from the probability function thereby optimizing the step of finding the nearest neighbor.

21. A process for use in answering queries comprising the steps of:

- a) clustering data stored in a database using a clustering technique to provide an estimate of the probability density function of the sample data;
- b) adding an additional column attribute to the database that represents the predicted cluster membership for each data record within the database; and
- c) rebuilding a data table of the database using the newly added column as an index to records in the table.

22. The process of claim 21 wherein an index for the data in the database is created on the additional column.

23. The process of claim 21 comprising the additional step of performing a nearest neighbor query to identify a nearest neighbor data point to a query data point.

24. The process of claim 22 wherein the nearest neighbor query is performed by finding a nearest cluster to the query data point.

25. The process of claim 24 additionally comprising the step of scanning data in a cluster identified as most likely to contain nearest neighbor based on a probability estimate for said cluster.

26. The process of claim 25 wherein if the probability that the nearest neighbor estimate is correct is above a certain threshold, the scanning is stopped, but if it is not, then scanning additional clusters to find the nearest neighbor.

27. In a computer data mining system, apparatus for evaluating data in a database comprising:

- a) one or more data storage devices for storing data records on a storage medium; the data records including data attributes; and

20

b) a computer having a rapid access memory and an interface to the storage devices for reading data from the storage medium and bringing the data records from the storage medium into the rapid access memory for subsequent evaluation;

c) the computer comprising a processing unit for evaluating at least some of the data records and for determining a probability density function for the records based on a clustering of data from data in the database into multiple numbers of data clusters, and said computer programmed to build an index for the data records in the database based on the probability density function, wherein said computer performs an approximate nearest neighbor analysis by choosing a specified cluster based on the index and evaluating records of the specified cluster for nearness to a given data record.

28. The apparatus of claim 27 wherein said computer stores data records having a common index based on cluster number in a file of records not part of database table.

29. The apparatus of claim 27 wherein the computer includes a database management component for setting up a database and using the index to organize data records in the database.

30. The apparatus of claim 29 wherein the computer builds an additional database of records for storage on the one or more data storage devices and wherein the data records of the additional database are augmented with a cluster attribute.

31. A computer-readable medium having computer-executable components comprising:

- a) a database component for interfacing with a database that stores data records made up of multiple data attributes;
- b) a modeling component for constructing and storing a clustering model that characterizes multiple data clusters wherein the modeling component constructs a model of data clustering that corresponds to a mixture of probability functions; and
- c) an analysis component for indexing the database on a cluster number for each record in the database wherein the indexing is performed based on a probability assessment of each record to the mixture of probability functions and for approximating a nearest neighbor query by determining an index for a sample record and scanning data records previously assigned a similar index.

32. The computer readable medium of claim 31 wherein said indexing component generates an augmented data record having a cluster number attribute for storage by the database component.

33. The computer readable medium of claim 31 wherein said modeling component is adapted to compare a new model to a previously constructed model to evaluate whether further of said data records should be moved from said database into said rapid access memory for modeling.

34. The computer readable medium of claim 31 wherein said modeling component is adapted to update said cluster model by calculating a weighted contribution by each of said data records in said rapid access memory.

21

35. A method for evaluating data records stored on a storage medium wherein each record has multiple data attributes that have been clustered to define a probability function of the data records stored on the storage medium; said method comprising the steps of evaluating the clusters of a clustering model and if the cluster separation between cluster centroids is of a sufficient size, assigning an index for each of the records on the storage medium based upon the probability function that is derived from the clustering model.

36. A method for evaluating data records stored on a storage medium wherein each record has multiple data attributes that have been characterized by a probability function found by clustering the data in the database to produce a clustering model; said method comprising the steps of assigning a cluster number for each of the records on the storage medium based upon the probability function, and finding a nearest neighbor from the data records for a query record by choosing between a scan of a subset of the database and a complete scan of the database based on the probability estimate generated by a statistical model of the data.

22

37. A method for performing an approximate nearest neighbor search of data records in a database stored on a storage medium wherein each record has multiple data attributes comprising:

- a) clustering the data records to define a cluster model having multiple clusters made up of probability functions to form a compressed characterization of the data records of the database;
- b) assigning clusters numbers as indexes to the data records based on the probability functions; and
- c) searching data records from a cluster to find a nearest neighbor within said cluster of a sample data record based on a nearness of the sample data record to the clusters that make up the cluster model.

38. The method of claim 37 wherein records from multiple numbers of clusters are searched to find a nearest neighbor to a sample data record.

39. The method of claim 38 wherein the number of clusters searched is based on a probability that an actual nearest neighbor is contained within an as yet unscanned cluster.

* * * * *